A proximity-based method to identify genomic regions correlated with a continuously varying environmental variable.

Evolutionary Bioinformatics

METHODOLOGY

A Proximity-Based Method to Identify Genomic Regions Correlated with a Continuously Varying Environmental Variable

Cornelia Di Gaetano[1,2], Giuseppe Matullo[1,2], Alberto Piazza[1,2], Moreno Ursino[3] and Mauro Gasparini[3]

[1]Department of Genetics, Biology and Biochemistry, University of Turin, Turin, Italy. [2]HuGeF, Human Genetics Foundation, Turin, Italy. [3]Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy.
Corresponding author email: cornelia.digaetano@unito.it

Abstract: Knowledge of markers in the human genome which show spatial patterns and display extreme correlation with different environmental determinants play an important role in understanding the factors which affect the biological evolution of our species. We used the genotype data of more than half a million single nucleotide polymorphisms (SNPs) from the data set Human Genome Diversity Panel (HGDP-CEPH -CEPH) and we calculated Spearman's correlation between absolute latitude and one of the two allele frequencies of each SNP. We selected SNPs with a correlation coefficient within the upper 1% tail of the distribution. We then used a criterion of proximity between significant variants to focus on DNA regions showing a continuous signal over a portion of the genome. Based on external information and genome annotations, we demonstrated that most regions with the strongest signals also have biological relevance. We believe this proximity requirement adds an edge to our novel method compared to the existing literature, highlighting several genes (for example *DTNB, DOT1L, TPCN2, RELN, MSRA, NRG3*) related to body size or shape, human height, hair color, and schizophrenia. Our approach can be applied generally to any measure of association between polymorphic frequencies and continuously varying environmental variables.

Keywords: adaptations, spatial patterns, latitude, point processes, outlier approach

Evolutionary Bioinformatics 2013:9 29–42

doi: 10.4137/EBO.S10211

This article is available from http://www.la-press.com.

Introduction

From an evolutionary point of view, human biological variation can result from natural selection, genetic drift and demographic processes. In human population genetics, several ways have been found to highlight genes that may be subject to selective pressures, and in recent years whole genome scanning techniques have made it possible to find signatures of selection.[1–4] The Human Genome Diversity Project (HGDP-CEPH) database[5] has been repeatedly investigated in order to identify markers in the human genome which show geographical patterns and to explain how different selective forces can shape human genetic variations across continents. One strategy for the detection of spatial selection signatures is the outlier approach.[2,6,7] Using genome-wide data sets genotyped in different human populations, genetic variables—such as single nucleotide polymorphisms (SNPs)—that exhibit extreme correlations with latitude or with other environmental determinants are identified as candidate targets for selective pressure. By "extreme correlation" we mean that the value of a certain statistic, measuring the strength of the relationship between allele frequencies and latitude or other environmental variables, falls in the tails of the distribution of the same statistic over the whole genome. Many choices are possible for the relevant statistic, ranging from a simple (either Pearson or Spearman) correlation coefficient between the latitude and the frequencies of either one or two alleles of a SNP to a Bayes factor comparing two models that do and do not, respectively, take into account the effect of a dichotomous environmental variable on the distribution of a genetic variant. From a technical point of view, the outlier approach is just a reformulation of the concept of *statistical significance*, ie, variation with respect to a reference distribution.

The outlier approach has been used to study sodium homeostasis balance as an example of adaptation. In hot and dry climates, genes influencing salt and water retention are favored by selection, explaining in this way large inter-ethnic differences in the prevalence of salt-sensitive hypertension.[8,9] Other important research has been conducted to assess the correlation between four variables that summarize climate and the frequencies of 873 tag SNPs in 82 genes related to energy metabolic pathways.[6] The outlier approach has also been used to demonstrate that allele frequencies of a subset of genes coding for blood group antigens vary with levels of pathogen richness, supporting the idea that these loci affect susceptibility to infectious diseases.[10] This finding, which is compatible with previous evidences on the correlation between HLA class I diversity and pathogen richness,[11] is important for stressing the role of diseases and pathogens, like virus protozoa fungi, in shaping human variations.[12] Finally, a very comprehensive article on the HGDP-CEPH database (enriched with the Hap Map and other human populations databases) has recently been published, in which the outlier approach is used to highlight polymorphisms and pathways correlated with ecoregion membership and diet.[13]

Our idea is to reinforce the outlier approach by considering a criterion of *proximity* between significant variants. In the search for targets of selective pressure, we believe it is important to focus on those DNA regions which *repeatedly* contain values which are labeled as significant by the outlier approach. In other words, we look for evidence of a continuous signal over a portion of the genome which can strengthen the significance of a cluster of markers labeled as significant by the outlier approach alone and we built statistical tools.

In this paper we therefore adopt a search-and-confirm approach which integrates the outlier approach by identifying regions of the genome where not just one, but a significant number of SNPs are located in the tails of the distribution of the relevant statistic, when compared to the number of SNPs originally genotyped in the same region. This is done in the following three steps, which are further illustrated in the complete workflow process diagram in Figure 1:

1. The outlier method: We identify 1% significant SNPs as having an absolute value of the Spearman correlation coefficient with latitude above its 99th percentile;
2. The proximity-based algorithm: Using the methods described in detail in the Materials and Methods section, we select candidate regions in the genome which exhibit the strongest signals, ie, the regions where the significant SNPs identified above are present at a significantly higher rate when compared to the number of originally genotyped SNPs;
3. Biological relevance: We investigate the biological relevance of the strongest signals by comparing our

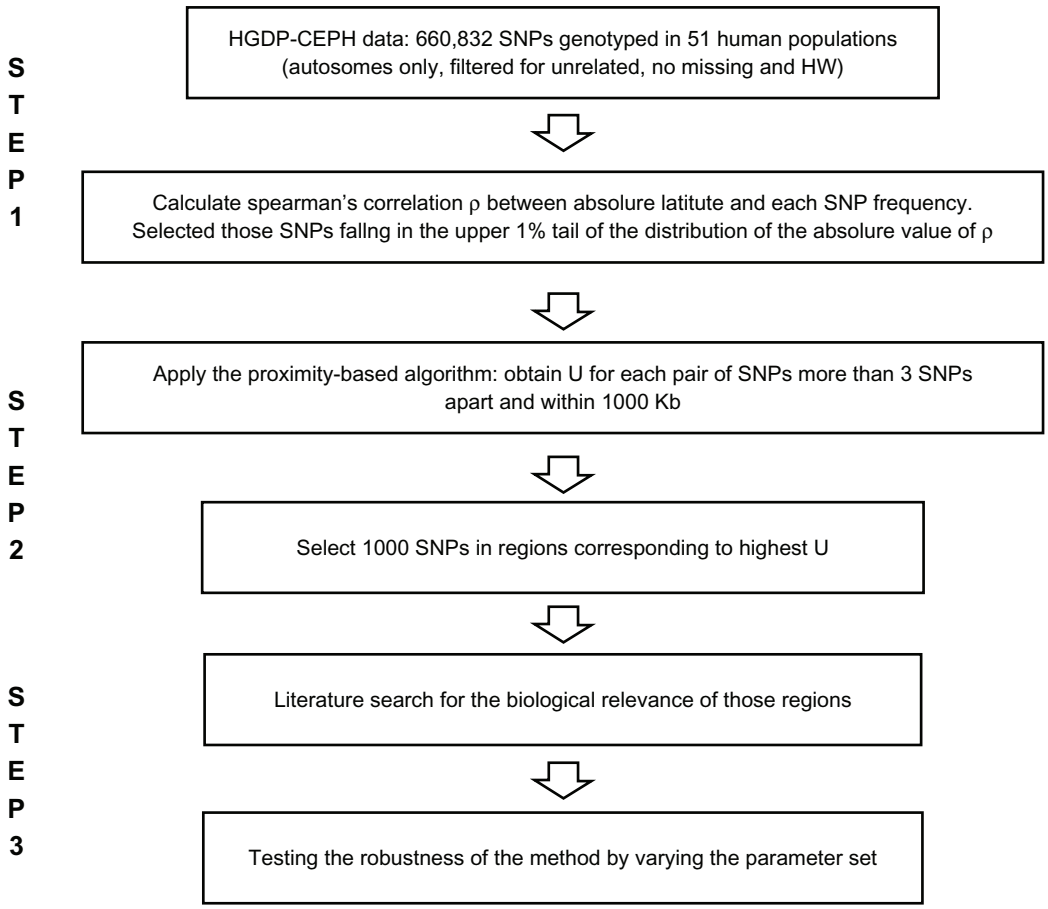

Figure 1. Graphical workflow process for the study.

data with results from Genome Wide Association studies (GWAs),[14] by studying the canonical pathway processes through gene-annotation enrichment analysis[15] and by comparing our analysis with previously published genomic scans for selective sweep.[3,16]

Materials and Methods

We describe here our methods with reference to the three-step process described in the Introduction.

Step 1: Our data and the outlier method

We used a data set of 660,832 SNPs genotyped in 51 human populations distributed worldwide from the HGDP-CEPH panel.[1] As underlined by a previous article,[17] within the HGDP-CEPH panel there are some closely-related individuals; in order to overcome this possible source of bias we excluded one member of each relative pair and we used 938 HGDP-CEPH individuals. Information about sample sizes and latitudes of the populations can be found on the CEPH homepage http://www.cephb.fr/en/hgdp/table.php.[5] Only 22 autosomes are included in our analysis; we also removed SNPs with more than 10% of missing genotypes and the ones that failed the Hardy-Weinberg *equilibrium* test in at least one population. After filtering, we use 545,209 SNPs.

Statistical analysis is performed using R.[18] We calculated Spearman's correlation (the correlation coefficient between the ranks of two variables) between absolute latitude and one of the two alleles of each SNP and, using the outlier approach, we identified those SNPs which have an absolute Spearman's correlation coefficient falling in the upper 1% tail of the distribution (Fig. 2).

Step 2: The proximity-based algorithm

For each chromosome, we now have two sequences of serial positions: one for all genotyped SNPs and one for the significant SNPs, the latter of which are included in the former. Each chromosome is indexed by the sequence of base pairs: as an approximation, we can view a chromosome as a linear segment and

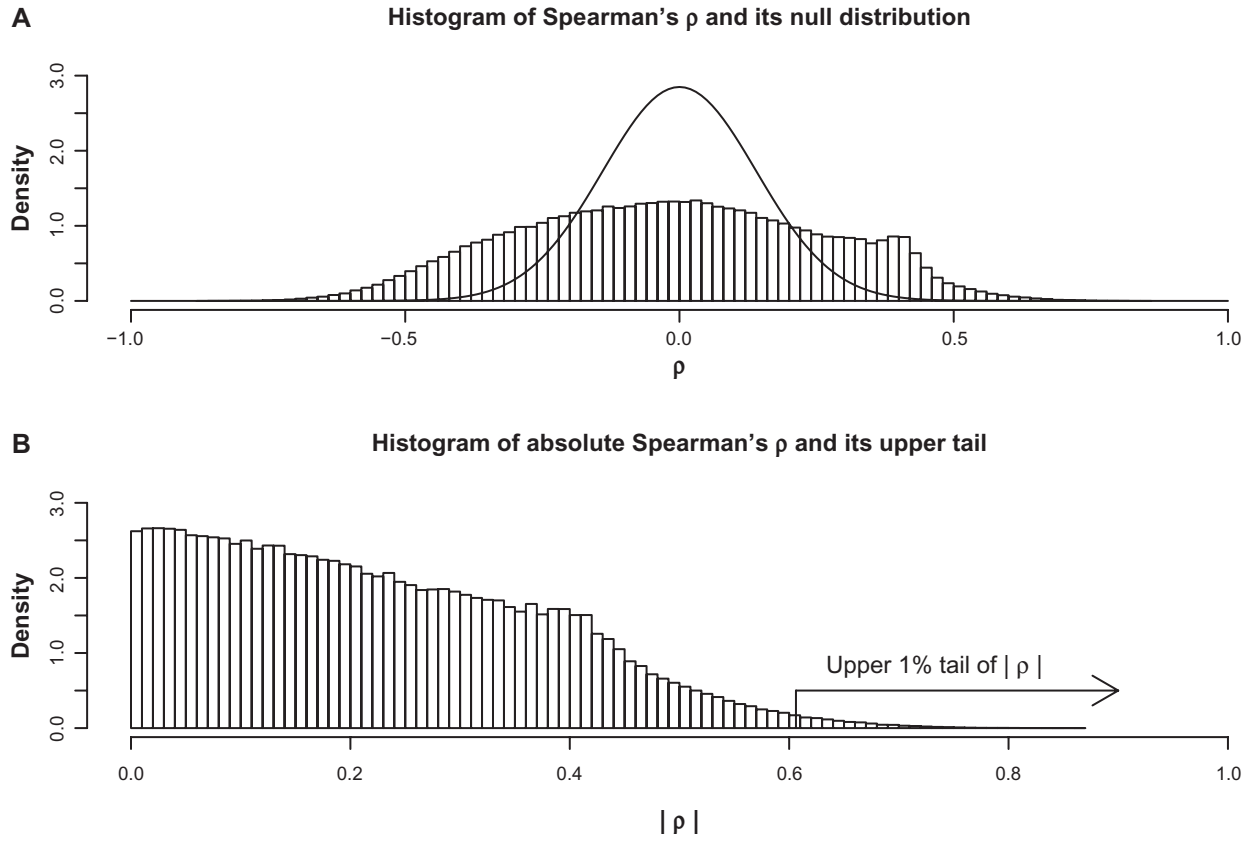

Figure 2. (**Panel A**) Histogram of the values of Spearman's correlation coefficient over all the SNPs and theoretical approximate density of the Spearman's correlation coefficient under the hypothesis of population null correlation. (**Panel B**) Histogram of the absolute values of Spearman's correlation coefficient over all the SNPs. Using the outlier approach, we identify significant SNPs in the 1% upper tail of this distribution.

the position of a SNP as a point of that linear segment. Based on the two sequences of points, we can define two cumulative counts depending on a generic point l, known in statistics as *counting processes*:

$S(l)$ = number of SNPs with a position smaller than or equal to l

$S_{.01}(l)$ = number of significant SNPs with a position smaller than or equal to l

with l varying from 1 (the first bp in the chromosome) to the position of the last bp of the chromosome. As an example, the two counting processes are plotted for chromosome 1 in Figure 3. Cumulative counts are a convenient way to compare the incidences of the different kinds of SNPs over different genomic regions (a simple dot plot would not do it, due to the sheer number of SNPs involved). If, over a certain segment of the chromosome, there is a greater-than-usual incidence of significant SNPs, then the relative increment of $S_{.01}(l)$ over that segment will be greater than the relative increment of $S(l)$ over the same segment. In other words, the graph of the $S_{.01}(l)$ counting process will be steeper than $S(l)$, up to a proportionality factor. Our proposal is to identify those genome regions which exhibit extreme concentrations of outlying SNPs.

We could formalize this search as a change-point problem for counting processes: in certain intervals to be estimated, the intensity of the $S(l)$ point process—a function modelling the instantaneous rate of incidence of the process—would be higher than in other regions. Due to the size of the problem and to the approximate nature of our search- and -confirm approach, we prefer a simpler *proximity-based algorithm* as follows.

For each pair of significant SNPs located at points l_1 and l_2, with $l_1 < l_2$ on the chromosome, we define

$$U(l_1, l_2) = \frac{S_{.01}(l_2) - S_{.01}(l_1)}{S(l_2) - S(l_1)}$$

ie, the observed incidence rate of significant SNPs per original SNP. This statistic over the sliding window

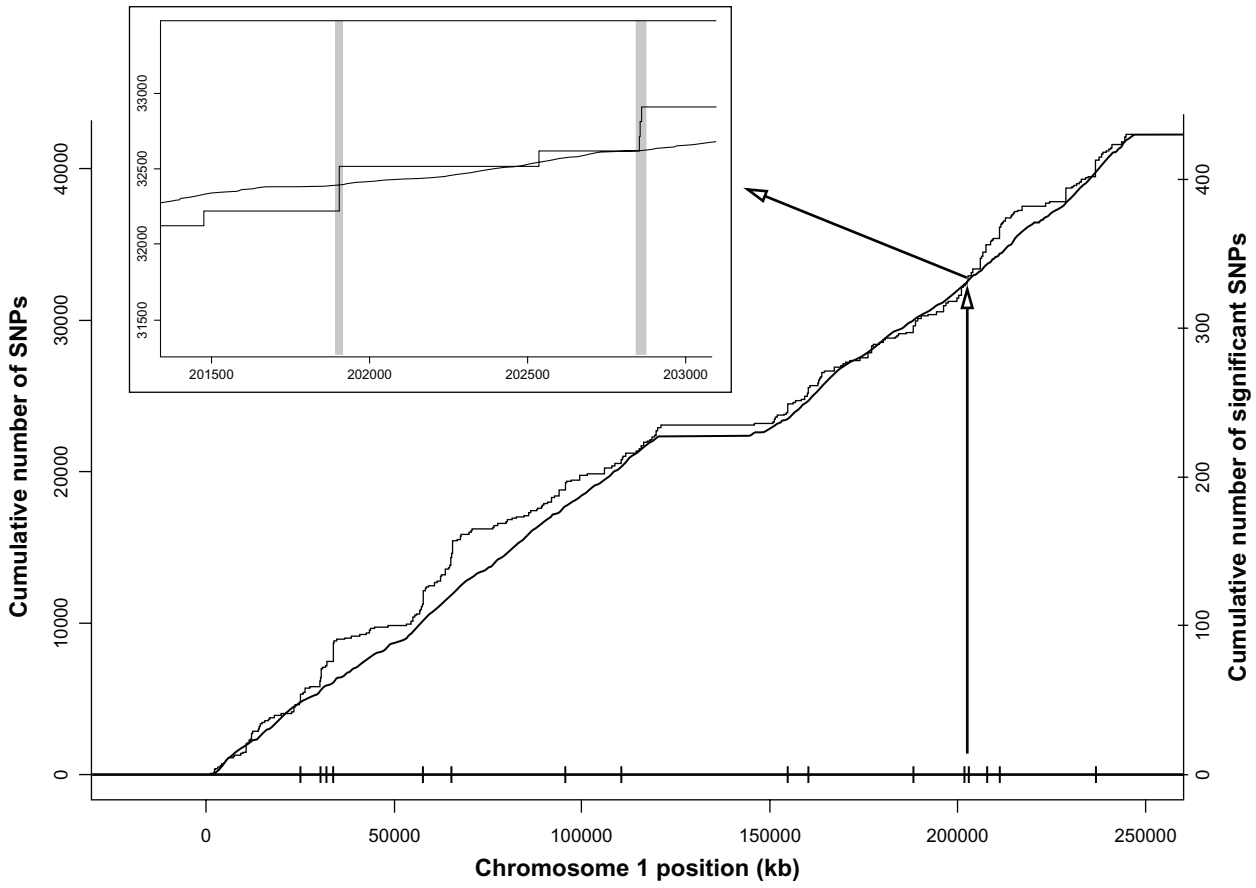

Figure 3. Counting process representation of the location of the candidate regions of chromosome 1.
Notes: The thicker step function represents cumulative counts of all originally genotyped SNPS and refers to the main ordinate scale, on the left. The thinner step function represents cumulative counts of significant SNPs and refers to the ordinate scale on the right. Sixteen regions identified by our method are shown as small vertical segments on the abscissa axis. The zooming box on the upper left part of the graph shows two of them (gray bands) located around position 202500 kb, as guided by the arrows.

(l_1, l_2) plays a central role in our proximity-based algorithm.

As a technical note, it would probably be a good idea to penalize large windows, for example by dividing the $U(l_1, l_2)$ statistic above by a penalty term $(l_2–l_1)^g$ with g equal to some number between 0 and 1. The final results would not change a lot (results not shown) and it would be difficult to commit to a specific g; therefore we decide to use the $U(l_1, l_2)$ statistic without a penalty term.

For each chromosome and for each significant SNP in position l_1, we computer $U(l_1, l_2)$ for each of the other significant SNPs in position l_2 within a distance of 1000 Kb from the original one. This is done to reduce the problem to a manageable size, under the assumption that relevant proximities are smaller than 1000 Kb.

We built the new reference distribution of all $U(l_1, l_2)$ values over all chromosomes, excluding from the analysis all $U(l_1, l_2)$ values relative to intervals (l_1, l_2) which included fewer SNPs than a threshold s, which has been chosen to be equal to 3 in this work. This is done to avoid very high automatic values of $U(l_1, l_2)$ when two significant SNPs happen to be adjacent. We selected the first 1000 SNPs contained in regions corresponding to the highest $U(l_1, l_2)$ values. A fixed number, rather than a fix tail area, was chosen to facilitate the discussion of the robustness of our method to varying parameters (see end of section Results).

Step 3: Biological relevance of the strongest signals

To accomplish step 3 as outlined in the Introduction, we proceeded to the biological cross-validation of our findings, which insofar had been based mainly on statistical grounds. We focused on the genes tagged by the SNPs we found, since our goal was to detect continuous signals coming from proximal groups of

SNPs belonging to the same gene. To link our findings to the results of genome wide data, we first compared our gene list with the June 2012 update of the Catalog of Publish ed GWAs.[14] Next, we scanned our gene list using a bioinformatic enrichment tool named Genecodis 2.0[15] to obtain a summary of the most enriched biological processes or pathways. Finally, we compared our analysis with previously published genomic scans for selective sweep in order to find possible overlaps in signals.

Results

We calculated Spearman's correlation between absolute latitude and one of the two alleles of the SNPs found in the HGDP-CEPH panel and, following the outlier approach, we identified those SNPs which have an absolute Spearman's correlation coefficient falling in the upper 1% tail of the distribution. The histogram of Spearman's correlations ρ's is plotted in Figure 2A. Its null distribution for 51 pairs of numbers has been overlaid on the same graph (Fig. 2A). It is a normal distribution with variance 1/50 due to a well-known result.[19] The discrepancy between the two distributions is due to SNPs which are correlated with latitude for reasons other than chance alone, for example due to environmental selection factors. Following the outlier approach, the upper 1% of the distribution of the absolute value of ρ, corresponding to $|\rho| > 0.606$, is identified in the histogram of the absolute value of ρ (Fig. 2B). It corresponds to 5452 outlying SNPs in the tails of the ρ distribution.

The candidate regions and the annotations emerging from the application of Step 2 described in the Introduction are contained in Additional 1 in the online supporting information. As an example, candidate regions which were identified in chromosome 1 are shown in Figure 3. The 1000 top SNPs emerging from the proximity-based algorithm enabled us to identify 467 intergenic and 533 genic SNPs, harboring 146 genes. We found 23 coding non synonymous (NS) changes and 6 coding synonymous changes. 372 were intronic and 107 were on the mRNA 3′UTR.

Finally, we gathered the biological knowledge of the strongest signals by comparing them to the Catalog of Published Genome-Wide Association Studies updated to June 2012. The genes which appear on this Catalog and additionally appear in candidate regions according to our proximity-based algorithm, are shown in Additional file 2. A short list of the most interesting signals are shown in Table 1. Several genes shown in that table are associated with metabolism-related phenotypes (like celiac disease for *IL21* interleukin 21, Gene id 59067)[20] and adiposity (*MSRA* Gene id4482) or variants associated with hair color in Europeans, like *TPCN2* gene (two pore segment channel 2, gene ID 219931)[21] and several with schizophrenia. At the same time, we compared our gene list with genes reported in OMIM. Several of our genes which show a correlation with latitude also implied some traits. For example, *DOT1L* gene (DOT1-like, histone H3 methyltransferase *Saccharomyces cerevisiae*) gene ID 84444 is associated with height[22] or *DTNB* gene dystrobrevin, beta ID 1838 which is affecting adult human height.[23] A complete table with the genes reported also in OMIM Disease database is in Additional file 3.

We analyzed Kyoto Encyclopedia of Genes and Genomes pathways (KEGG) using as reference set all genes in the Entrez-gene database and, as a statistical test, the hypergeometric one with a Benjamini-Hochberg correction for multiple testing at significance level equal to 0.05. Several KEGG pathways reached significance. The first was the extracellular matrix (ECM) receptor interaction (KEGG number: hsa04512) for the following genes: *RELN* reelin gene ID 5649; *ITGB6* integrin beta 6 gene ID 3694; *COL6A3* collagen, type VI, alpha 3 gene ID 1293. This pathway reaches a raw *P*-value of the hypergeometric test equal to 0.0011 and a *P*-value adjusted for multiplicity around 0.01. In order to look for overlaps with scans of the human genome for signals of positive natural selection, we compared our results with SNPs with significant composite of multiple signals (CMS) but only one intersection was found between the two gene lists concerning rs2256670 and rs2711853 both on *RELN* reelin, gene ID 5649.[16] A variety of choices were made in the actual implementation of the proximity-based algorithm described in Step 2 in the previous section. The two most important parameters set to reasonable values are (a) the maximum distance over which we search, which is set to 1000 Kb in Step 2, and (b) the minimum number of consecutive SNPs required, which is set to 3 in Step 2. In order to study the robustness of our method with respect to different values of these parameters, we varied the maximum distance and

Table 1. List of several genes reported in previously published GWAs and showing continuous correlation signals with our proximity based method.

Reported gene(s)	Trait	Region	NCBI ID	Gene description	Reference
C9orf3	Erectile dysfunction and prostate cancer treatment	9q22.32	84909	Chromosome 9 open reading frame 3	41
ABL1	Response to amphetamine	9q34.12	25	v-abl Abelson murine leukemia viral oncogene homolog 1	42
DTNB	Adult human height	2p23.3	1838	Dystrobrevin, beta	23
DTNB	Coronary heart disease	2p23.3	1838	Dystrobrevin, beta	43
TPCN2	Hair pigmentation in Europeans	11q13.3	219931	Two pore segment channel 2	21
DOT1L	Associated with height	19p13.3	84444	DOT1-like, histone H3 methyltransferase (*S. cerevisiae*)	22,35
RELN	Susceptibility and clinical phenotype in multiple sclerosis	7q22.1	5649	Reelin	44
RELN	Increases the risk of schizophrenia only in women	7q22.1	5649	Reelin	34
IL21	Celiac disease	4q27	59067	Interleukin 21	38,45
DOCK2	Protein quantitative trait loci	5q35.1	1794	Dedicator of cytokinesis 2	46
FRMD4B	Celiac disease	3p14.1	23150	FERM domain containing 4B	38
MAGI2	Hippocampal atrophy	7q21.11	9863	Membrane associated guanylate kinase, WW and PDZ domain containing 2	47
NCALD	Cognitive performance	8q22.3	83988	Neurocalcin delta	48
NRG3	Response to iloperidone treatment (QT prolongation)	10q23.1	10718	Neuregulin 3	49
RUNX3	Celiac disease	1p36.11	864	Runt-related transcription factor 3	38
SDK1	Quantitative traits	7p22.2	221935	Sidekick homolog 1 (chicken)	50
MSRA	Adiposity	8p23.1	4482	Methionine sulfoxide reductase A	27
MSRA	Hypertension	8p23.1	4482	Methionine sulfoxide reductase A	28
MSRA	Schizophrenia	8p23.1	4482	Methionine sulfoxide reductase A	51
MSRA	Bipolar disorder and schizophrenia	8p23.1	4482	Methionine sulfoxide reductase A	26

noticed (not shown) that the results where unchanged for distances down to 100 Kb.

The algorithm is instead sensitive to the minimum number of consecutive SNPs required: if we increase it from 3 to 5, for example (it would not make sense to consider a minimum much higher than 5), different SNPs and regions turn out to be significant, as shown in Table 2. For example, the number of selected SNPs shared when applying a minimum of 5 and when applying a minimum of 3 is 64%. This made us consider what would happen for varying this threshold. The changes are not dramatic (Table 2) but some interesting genes, like *AGT, ADCY9* and *WWOX* would come out from the analysis with a threshold equal to 5.

Discussion

In this paper we examined the HGDP-CEPH data again by integrating the outlier approach with a novel proximity-based algorithm.

Only latitude was used for ecological conditions, rather than using a multiplicity of variables as in Hancock et al[6] for example. We made this choice for the sake of simplicity, since latitude is correlated with different variables like short wave radiation flux, mean winter and summer temperatures, rainfall and pathogen richness. It should therefore provide a good proxy for the selective pressures that shaped variation in our genome. Even though we use a simple correlation measure such as Spearman's ρ with latitude only, we emphasize that the resulting signal should be a continuous and persistent proportion of background information, represented by all originally genotyped SNPs. We believe this proximity requirement adds an edge to our novel method when compared to existing literature. Our approach is applicable to any measure of association between polymorphic frequencies and environmental variables. It could be applied, for example, to complex statistics such as the minimum rank statistic, based on Bayes Factors and on rank transformations, of Hancock et al.[13]

Table 2. Percentage of common SNPs when varying the minimum number of consecutive SNPs required.

% concordance	3 SNPs	4 SNPs	5 SNPs
3 SNPs	100.00%	74.70%	64.00%
4 SNPs		100.00%	80.80%
5 SNPs			100.00%

With our method we identifed different genes, some of them already reported in the literature, dealing with different traits or diseases. GWAs include the scanning of all or most of the genes of different individuals aimed at finding susceptibility loci for traits or diseases. GWAs, so far, have allowed the identification of more than 7688 associated SNPs in humans. We compared our list of genes with GWAs results. Some interesting signals can be pointed out, for instance the correlation between skin pigmentation and latitude. It is well known that two coding variants in *TPCN2* are associated with hair color in Europeans.[21] At the same time *MSRA* (methionine sulfoxide reductase A gene) is related to the melanin formation in the hair follicle melanocyte.[24] Remarkably, *MSRA* gene is also related to schizophrenia[25,26] but also with adiposity[27] and hypertension.[28]

Several other genes in our list (see Additional file 1) can be associated with vitamin D related genes, known to show a latitude driven cline.[7] An example is *SMARCA2*, (SWI/SNF related, matrix associated, act in dependent regulator of chromatin, subfamily a, member 2), described as a component of a human multiprotein complex that that interacts directly with the vitamin D receptor. Schizophrenia genes are correlated with latitude and in our list several schizophrenia genes appear, like *GRID1*,[29,30] *MAGI2,*[31] *NRG3,*[32] *NRXN3,*[33] *RARB* and *RELN*.[34]

Region *CYP19A1* in our list is known from GWAs to exhib it association with adult height[35,36] whose distribution is related to latitude. Two more genes in our list, DOT1-like, histone H3 methyltransferase (*S. cerevisiae*)[22,35] and dystrobrevin, beta[23] are reported in OMIM to be related with height.

Several others genes are related to Celiac Disease (CD) which strongly correlates with latitude. Infectious agents are implicated in the pathogenesis of many autoimmune diseases like CD. This observation may imply that there is a relationship between one or more infectious agents, latitude related environmental exposure to gluten and others genetic susceptibility loci, and the development of this disease. For a complete review see Plot and Amital, 2009.[37] The *RUNX3* gene and *IL21*, in our list, are implicated with CD.[38] In the same paper, another gene *FRMD4B* previously known as *GRSP1*, appearing in our Table 1 is also associated with CD.[38] *RUNX3* gene is also required for CD8 T cell development during thymopoiesis.[39]

One of the most interesting genes highlighted by our work is *ANK2* (ankyrin 2, neuronal) which is implicated in cardiac arrhythmias due to abnormal variations in QT interval.[40]

Finally, the enrichment of genes in the KEGG pathway called extracellular matrix (ECM) receptor interaction (KEGG number: hsa04512) is note worth because these molecules are exploited by a number of pathogenic micro-organisms as receptors for cell entry. This can be interpreted as a signal of different forces played by pathogens on living cells in different environments.

Conclusions

Our study complements the growing body of knowledge surrounding scans for natural selection in humans using a method that uses the proximity criterion in addition to the outlier approach. Our findings support the hypothesis that latitudinal genetic diversity gradients are present in humans and reflect genetic adaptations to different environmental pressures that have shaped the human genome.

Acknowledgements

The authors thank the Human Genetic Foundation (HuGeF) Laboratory of Genomic Variation in Human Populations and Complex Disease group for biological and bioinformatics discussion.

Authors' Contributions

CDG designed the biological rationale for the study and MG provided the statistical tools to implement it, both authors wrote and revised the manuscript. MU and AP performed the data analysis. GM revised the manuscript. All authors read and approved the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. Feb 22, 2008; 319(5866):1100–4.
2. Coop G, Pickrell JK, Novembre J, et al. The role of geography in human adaptation. *PloS Genet.* 2009;5(6):e1000500.
3. Pickrell JK, Coop G, Novembre J, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research.* 2009; 19(5):826–37.
4. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 2010;20(4): R208–15.
5. Cann HM, de Toma C, Cazes L, et al. A human genome diversity cell line panel. *Science.* 2002;296(5566):261–2.
6. Hancock AM, Witonsky DB, Gordon AS, et al. Adaptations to climate in candidate genes for common metabolic disorders. *PloS Genet.* 2008; 4(2):e32.
7. Amato R, Pinelli M, Monticelli A, Miele G, Cocozza S. Schizophrenia and vitamin D related genes could have been subject to latitude-driven adaptation. *BMC Evol Biol.* 2010;10:351.
8. Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A. CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet*. Dec 2004;75(6):1059–69. Epub Oct 18, 2004.
9. Young JH, Chang YP, Kim JD, et al. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PloS Genet.* 2005;1(6):e82.
10. Young JH, Chang YP, Kim JD, et al. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PloS Genet.* 2005;1(6):e82.
11. Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 2005;15(11):1022–7.
12. Pozzoli U, Fumagalli M, Cagliani R, et al. The role of protozoa-driven selection in shaping human genetic variability. *Trends Genet*. 2010;26(3):95–9.
13. Hancock AM, Witonsky DB, Ehler E, et al. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A*. 2010;107 Suppl 2:8924–30.
14. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106(23):9362–7.
15. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*. 2007;8(1):R3.
16. Grossman SR, Shylakhter I, Karlsson EK, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science.* 2010;327(5967):883–6.
17. Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet.* 2006;70(Pt 6):841–7.
18. Development Core Team R. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

19. Hotelling HPR. Rank correlation and tests of significance involving no assumption of normality. *Ann Math Statist.* 1936;7(14).
20. Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet.* 2008; 40(4):395–402.
21. Sulem P, Gudbjartsson DF, Stacey SN, et al. Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet*. 2008;40(7): 835–7.
22. Lettre G, Jackson AU, Gieger C, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet.* 2008;40(5):584–91.
23. Gudbjartsson DF, Walters GB, Thorleifsson G, et al. Many sequence variants affecting diversity of adult human height. *Nat Genet*. 2008;40(5): 609–15.
24. Schallreuter KU, Salem MM, Hasse S, Rokos H. The redox—biochemistry of human hair pigmentation. *Pigment Cell Melanoma Res.* 2011;24(1): 51–62.
25. Walss-Bass C, Soto-Bernardini MC, Johnson-Pais T, et al. Methionine sulfoxide reductase: a novel schizophrenia candidate gene. *Am J Med Genet B Neuropsychiatr Genet*. 2009;150B(2):219–25.
26. Bergen SE, O'Dushlaine CT, Ripke S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry.* 2012;17(9):880–6.
27. Lindgren CM, Heid IM, Randall JC, et al. Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PloS Genet.* 2009;5(6):e1000508.
28. Levy D, Larson MG, Benjamin EJ, et al. Framingham Heart Study 100 K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Med Genet.* 2007;8 Suppl 1:S3.
29. Treutlein J, Muhleisen TW, Frank J, et al. Dissection of phenotype reveals possible association between schizophrenia and Glutamate Receptor Delta 1 (GRID1) gene promoter. *Schizophr Res.* 2009;111(1–3):123–30.
30. Chen X, Lee G, Maher BS, et al. GWA study data mining and independent replication identify cardiomyopathy-associated 5 (CMYA5) as a risk gene for schizophrenia. *Mol Psychiatry.* 2011;16(11):1117–29.
31. Koide T, Banno M, Aleksic B, et al. Common variants in MAGI2 gene are associated with increased risk for cognitive impairment in schizophrenic patients. *PLoS One.* 2012;7(5):e36836.
32. Kao WT, Wang Y, Kleinman JE, et al. Common genetic variation in Neuregulin 3 (NRG3) influences risk for schizophrenia and impacts NRG3 expression in human brain. *Proc Natl Acad Sci U S A.* 2010;107(35): 15619–24.
33. Gauthier J, Siddiqui TJ, Huashan P, et al. Truncating mutations in NRXN2 and NRXN1 in autism spectrum disorders and schizophrenia. *Hum Genet.* 2011;130(4):563–73.
34. Shifman S, Johannesson M, Bronstein M, et al. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. *PloS Genet.* 2008;4(2):e28.
35. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature.* 2010;467(7317):832–8.
36. Okada Y, Kamatani Y, Takahashi A, et al. A genome-wide association study in 19 633 Japanese subjects identified LHX3-QSOX2 and IGF1 as adult height loci. *Hum Mol Genet.* 2010;19(11):2303–12.
37. Plot L, Amital H. Infectious associations of Celiac disease. *Autoimmun Rev.* 2009;8(4):316–9.
38. Dubois PC, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet.* 2010;42(4): 295–302.
39. Woolf E, Xiao C, Fainaru O, et al. Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. *Proc Natl Acad Sci U S A.* 2003;100(13):7731–6.
40. Sedlacek K, Stark K, Cunha SR, et al. Common genetic variants in ANK2 modulate QT interval: results from the KORA study. *Circ Cardiovasc Genet.* 2008;1(2):93–9.
41. Kerns SL, Ostrer H, Stock R, et al. Genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with the development of erectile dysfunction in African-American men after radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys.* 2010;78(5):1292–300.
42. Hart AB, Engelhardt BE, Wardle MC, et al. Genome-wide association study of d-amphetamine response in healthy volunteers identifies putative associations, including cadherin 13 (CDH13). *PLoS One.* 2012;7(8):e42646.
43. Lettre G, Palmer CD, Young T, et al. Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. *Plos Genet.* 2011;7(2):e1001300.
44. Baranzini SE, Wang J, Gibson RA, et al. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet.* 2009;18(4):767–78.
45. Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet.* 2008; 40(4):395–402.
46. Melzer D, Perry JR, Hernandez D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PloS Genet.* 2008; 4(5):e1000072.
47. Potkin SG, Guffanti G, Lakatos A, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One.* 2009;4(8):e6501.
48. Cirulli ET, Kasperaviciute D, Attix DK, et al. Common genetic variation and performance on standardized cognitive tests. *Eur J Hum Genet.* 2010;18(7):815–20.
49. Volpi S, Heaton C, Mack K, et al. Whole genome association study identifies polymorphisms associated with QT prolongation during iloperidone treatment of schizophrenia. *Mol Psychiatry.* 2009;14(11):1024–31.
50. Lowe JK, Maller JB, Pe'er I, et al. Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PloS Genet.* 2009;5(2):e1000365.
51. Ma X, Deng W, Liu X, et al. A genome-wide association study for quantitative traits in schizophrenia in China. *Genes Brain Behav.* 2011;10(7):734–9.

```
#########################################################################
#
# proximity.R – by MU and MG, November 2012
#               R programs to analyze the HGDP-CEPH data according to
#               the proximity-based method in Di Gaetano et al.
#
#########################################################################

# the following instructions assume that data have been read from the
# text files from the web page http://hagsc.org/hgdp/files.html
#  via,  for  example,  read.table("HGDP_Map.txt")  or  read.csv2("id2.csv",
sep = ",")
# into a dataframe called "data"

### function to compute allele's frequencies

freqfun<- function(data){
k<-0
freq <- namefreq <- NULL

for (i in dimnames(data)[[2]][-(1:3)]){

### do the frequency table
tav <- table(data[,3],data[,i])
### we exclude the tables with dim 1 or 2, in which
### there is no variation

### heterozygous, all homozygous and missing
if (dim(tav)[[2]]==4) {
# search of allele with greatest frequency
allmagg <- c(sum(tav[,2]),sum(tav[,4]))
if (allmagg[1] > allmagg[2]) magg <- 2 else magg <-4
freq <- cbind(freq, round((tav[,magg]+tav[,3]/2)/(tav[,2]+tav[,3]+tav[,4]),3))
# use the variable name for the table
namefreq <- c(namefreq,i)
}

### heterozygous, all homozygous and no missing
if (dim(tav)[[2]]==3 & (dimnames(tav)[[2]][1]!="--")) {
# search of allele with greatest frequency
allmagg <- c(sum(tav[,1]),sum(tav[,3]))
if (allmagg[1] > allmagg[2]) magg <- 1 else magg <-3

freq <- cbind(freq, round((tav[,magg]+tav[,2]/2)/(tav[,1]+tav[,2]+tav[,3]),3))
# use the variable name for the table
namefreq <- c(namefreq,i)
}
```

```
dimnames(freq)[[2]] <- namefreq
cat(k <- k+1,"\n")

}
freq
}

n_min = 5
finestra = 1000000

### Computation of U for chrom 1

ovr1_sig <- ovr1[ovr1[,"Lsig01"] == "sig01",]
matrice_rappovr1 <- NULL

# loop over all significant SNPs

for (i in 1:(dim(ovr1_sig)[1]-1) ) { #last SNP is automatically processed

# p = number of subsequent snps to that processed
p <- n_min-1
while ( (i+p)<= dim(ovr1_sig)[1] & (ovr1_sig[i+p,4]-ovr1_sig[i,4])<= finestra )
{
   iniz <- ovr1_sig[i,4]
   fin <- ovr1_sig[i+p,4]
   u <- (ovr1_sig[i+p,9]-ovr1_sig[i,9]+1)/(ovr1_sig[i+p,7]-ovr1_sig[i,7]+1)
   x <- c(i,p+1,1,iniz,fin,u)
   matrice_rappovr1 <- rbind(matrice_rappovr1,x)
   p <- p+1
}

}
dimnames(matrice_rappovr1)[[2]] <- c("SNP","n SNP", "chrom", "reg in", "reg
fin","U")

### Selection of SNP
N = 1000

# union of results of all chromosomes
matrice_totale = rbind(matrice_rappovr1,matrice_rappovr2,matrice_rappovr3,
               matrice_rappovr4,matrice_rappovr5,matrice_rappovr6,
matrice_rappovr7,

     matrice_rappovr8,matrice_rappovr9,matrice_rappovr10,matrice_rappovr11,

     matrice_rappovr12,matrice_rappovr13,matrice_rappovr14,matrice_rappovr15,
```

```
     matrice_rappovr16,matrice_rappovr17,matrice_rappovr18,matrice_rappovr19,
        matrice_rappovr20,matrice_rappovr21,matrice_rappovr22)

matrice_totale <- data.frame(matrice_totale)
# decreasing order
matrice_totale <- matrice_totale[order(matrice_totale$U,decreasing = TRUE),]

# loop to extract result
ovr_sig <- list(ovr1_sig,ovr2_sig,ovr3_sig,ovr4_sig,ovr5_sig,ovr6_sig,ovr7_sig,

     ovr8_sig,ovr9_sig,ovr10_sig,ovr11_sig,ovr12_sig,ovr13_sig,ovr14_sig,
             ovr15_sig,ovr16_sig,ovr17_sig,ovr18_sig,ovr19_sig,ovr20_sig,
             ovr21_sig,ovr22_sig)
z = 1
selezione <- NULL
n_SNP = 0

while (n_SNP < N )
        {
        a = matrice_totale[z,1]
        b = matrice_totale[z,2]
        c = matrice_totale[z,3]

        sel <- cbind( ovr_sig[[c]][a:(a+b-1),1], rep(c,b) )
        selezione <- rbind(selezione, sel)

        z <- z+1

        # test to obtain unique solutions
        selezione <- unique(selezione)

        n_SNP <- dim(selezione)[1]
}

dimnames(selezione)[[2]] <- c("name_SNP" , "chrom" )
selezione <- selezione[1:N,]
```

Additional Files

Additional file 1: SNPs and regions from the proximity-based algorithm.

Additional file 1 in the online supporting information contains all regions selected by the proximity-based method, duly annotated.

Additional file 2: The complete list of genes reported in previously published GWAs and showing continuous correlation signals with our proximity based method.

Additional file 3: The complete list of genes reported in OMIM and showing continuous correlation signals with our proximity based method.

Additional file 4: R scripts.

Additional file 4 in the online supporting information contains R scripts to perform the necessary calculations.